DOCUMENT RESUME

ED 418 154 TM 028 252

AUTHOR	Perry, Nancy E.; Meisels, Samuel J.
TITLE	How Accurate Are Teacher Judgments of Students' Academic Performance? Working Paper Series.
INSTITUTION	National Opinion Research Center, Chicago, IL.
SPONS AGENCY	National Center for Education Statistics (ED), Washington, DC.
REPORT NO	NCES-WP-96-08
PUB DATE	1996-04-00
NOTE	49p.
CONTRACT	RN94094001
AVAILABLE FROM	U.S. Department of Education, Office of Educational Research and Improvement, National Center for Education Statistics, 555 New Jersey Avenue, N.W., Room 400, Washington, DC 20208-5652.
PUB TYPE	Reports - Evaluative (142)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	*Academic Achievement; Educational Assessment; Elementary Education; Literature Reviews; Longitudinal Studies; National Surveys; *Research Design; *Student Evaluation; *Teacher Attitudes; Teacher Expectations of Students; *Test Construction
IDENTIFIERS	Accuracy; *Early Childhood Longitudinal Study

ABSTRACT

The Early Childhood Longitudinal Study (ECLS) is a study that focuses on children's early school experiences beginning with kindergarten. The ECLS was developed by the National Center for Education Statistics. Approximately 23,000 children were selected as they entered kindergarten, and were followed as they moved through fifth grade. This paper is one of several that were prepared in support of ECLS design efforts. Literature documenting the accuracy of teachers' judgments of students' academic performance is reviewed in this paper. The first section examines the methods by which teacher judgments have been evaluated and summarizes findings from this research. Then research findings that relate specifically to the technical adequacy of these measures are examined. Implications are drawn from this research for using teacher judgment measures in the ECLS. Finally, recommendations are made for designing teacher assessment measures that are trustworthy and will serve the purposes of the ECLS adequately. Overall, the research indicates that teachers can make accurate judgments of students' academic performance, and some investigators have shown that teachers' judgments are even better predictors of students' future performance than standardized measures. Findings also suggest that teachers' judgments can be valid and reliable if certain precautions are taken. In general, it appears that more direct measures yield more accurate and meaningful judgments about students' academic achievements. Ways to make more indirect measures more accurate are discussed. (Contains 39 references.) (SLD)

ED 418 154

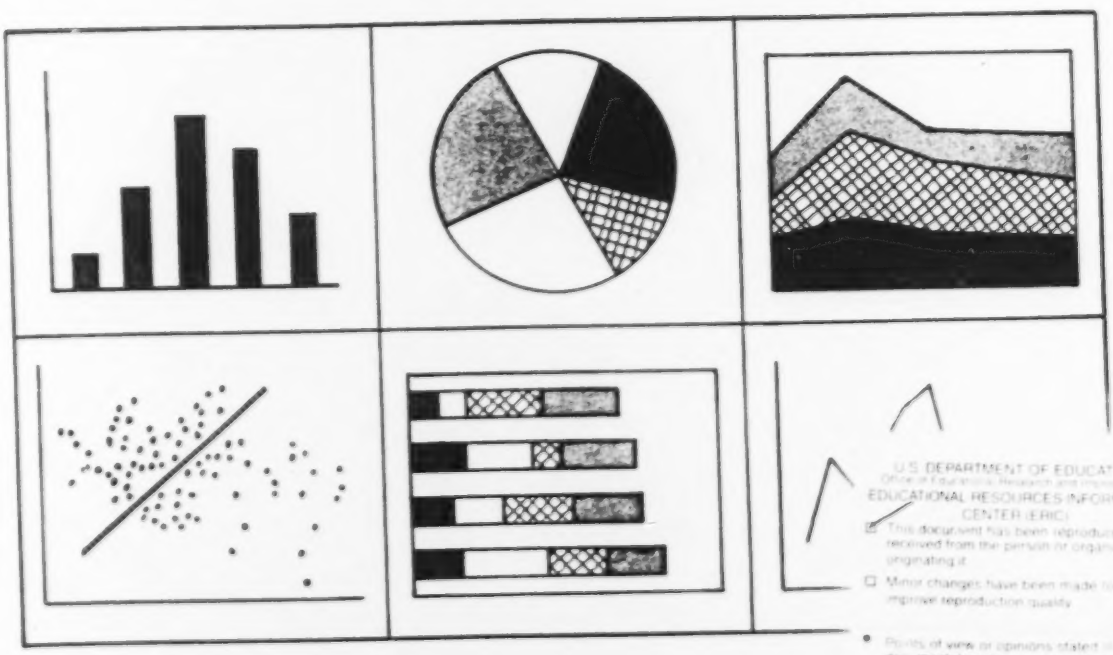
NATIONAL CENTER FOR EDUCATION STATISTICS

Working Paper Series

How Accurate are Teacher Judgments of Students' Academic Performance?

Working Paper No. 96-08

April 1996



U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

TM028252

U.S. Department of Education
Office of Educational Research and Improvement

How Accurate are Teacher Judgments of Students' Academic Performance?

Working Paper No. 96-08

April 1996

Contact: Jerry West
ECLS Project Officer
(202) 219-1574

U.S. Department of Education
Richard W. Riley
Secretary

Office of Educational Research and Improvement
Sharon P. Robinson
Assistant Secretary

National Center for Education Statistics
Jeanne E. Griffith
Acting Commissioner

Data Development and Longitudinal Studies Group
John Ralph
Acting Associate Commissioner

National Center for Education Statistics

The purpose of the Center is to collect and report "statistics and information showing the condition and progress of education in the United States and other nations in order to promote and accelerate the improvement of American education."—Section 402(b) of the National Education Statistics Act of 1994 (20 U.S.C. 9001).

April 1996

Foreword

Each year a large number of written documents are generated by NCES staff and individuals commissioned by NCES which provide preliminary analyses of survey results and address technical, methodological, and evaluation issues. Even though they are not formally published, these documents reflect a tremendous amount of unique expertise, knowledge, and experience.

The ***Working Paper Series*** was created in order to preserve the information contained in these documents and to promote the sharing of valuable work experience and knowledge. However, these documents were prepared under different formats and did not undergo vigorous NCES publication review and editing prior to their inclusion in the series. Consequently, we encourage users of the series to consult the individual authors for citations.

To receive information about submitting manuscripts or obtaining copies of the series, please contact Suellen Mauchamer at (202) 219-1828 or U.S. Department of Education, Office of Educational Research and Improvement, National Center for Education Statistics, 555 New Jersey Ave., N.W., Room 400, Washington, D.C. 20208-5652.

Susan Ahmed
Chief Mathematical Statistician
Statistical Standards and
Services Group

Samuel S. Peng
Director
Methodology, Training, and Customer
Service Program

How Accurate are Teacher Judgments of Students' Academic Performance?

Nancy E. Perry
Simon Fraser University

Samuel J. Meisels
The University of Michigan

April 1996

Prepared for the National Center for Education Statistics under contract RN94094001 with the National Opinion Research Center. The views expressed are those of the authors; no endorsement by the government should be inferred. The authors are indebted to Julie Nicholson and Sally Atkins-Burnett for their contributions to earlier versions of this paper.

Preface

The **Early Childhood Longitudinal Study (ECLS)** is a study that will focus on children's early school experiences beginning with kindergarten. The ECLS is being developed under the sponsorship of the U.S. Department of Education, National Center for Education Statistics (NCES), with additional financial and technical support provided by the Administration of Children, Youth and Families, U.S. Department of Education's Office of Special Education Programs and Office of Indian Education, and the U.S. Department of Agriculture's Food and Consumer Service. Approximately 23,000 children throughout the country will be selected to participate as they enter kindergarten and will be followed as they move from kindergarten through 5th grade. Base-year data will be collected in the fall of 1998, with additional spring follow-up data collections scheduled for 1999 through 2004. Information about children's neighborhoods, families, schools, and classrooms will be collected from parents, teachers, and school administrators.

Because of the magnitude and complexity of the ECLS, NCES has set aside an extended period of time for planning, designing, and testing the instruments and procedures that will be used in the main study. NCES and its contractor, the National Opinion Research Center, are using this time to examine a variety of issues pertaining to the sampling and assessment of young children and their environments. The design phase of the study will culminate in a large-scale field test during the 1996-97 school year.

NCES has sought the participation and input of many individuals and organizations throughout the design phase of the ECLS. The participation of these individuals and organizations has resulted in a set of design papers that identify policy and research questions in early education, map the content of the ECLS study instruments to these questions, explore and evaluate different methods for assessing the development of children and for capturing data about their homes, schools, and classrooms.

This paper is one of several that were prepared in support of ECLS design efforts. While the information and recommendations found in this paper have contributed to the design of the ECLS, specific methods and procedures may or may not actually be incorporated into the final ECLS design. It is our hope that the information found in this paper not only will provide background for the development of the ECLS, but will be useful to researchers developing studies of young children and their education experiences.

Jerry West
ECLS Project Officer

Jeffrey A. Owings
Program Director
Data Development and Longitudinal Studies Group

Table of Contents

Foreword iii

Preface v

How Accurate Are Teacher Judgments of Students' Academic Performance? 2

Evaluating Teachers' Judgments 4
Methods 4
Direct vs. indirect measures 4
Judgment specificity 8
Norm-referenced vs. criterion-referenced judgments 9
Summary 11

The Technical Adequacy of Teachers' Judgments 11
Validity 11
Criterion-related validity 12
Construct validity 14
Content validity 15
Summary 16
Reliability 17
Consistency within teachers 17
Consistency across raters 17
Internal Consistency 19
Summary 19
Equity 20
Summary 26

Limitations of Research on Teachers' Judgments 26

Summary and Implications 28

Recommendations 30

References 33

How Accurate Are Teacher Judgments of Students' Academic Performance?

Assessments that rely on teacher judgments of students' academic performance are used widely in both research and applied settings. In research settings, they contribute to evaluations of intervention studies, classroom processes, and children's intellectual, socio-emotional, and behavioral development (Hoge, 1985). In applied settings, teachers rely at least as often on their own judgments as they do on more objective measures in evaluating students' achievements, planning instruction, and reporting to parents (Sharpley & Edgar, 1986; Stiggins, 1987). Teachers' judgments are also used for screening and diagnostic decisions about referrals and special placements for individual students (Hoge, 1985). Finally, district and state level assessments are making increasingly greater use of teacher observation and judgment as a means of evaluating students' performances in such areas as writing, science, and visual or performing arts (Stiggins, 1987).

Some researchers argue that teachers can be valid assessors of their students. They claim that since teachers observe and interact with students on a daily basis, they are in the best position to evaluate their students' intellectual, socio-emotional, and behavioral accomplishments (Calfee & Hiebert, 1991; Hopkins, George, & Williams, 1985; Kenny & Chekaluk, 1993). Other researchers express concerns about the trustworthiness (i.e., validity and reliability) of these assessments (Hoge & Coladarci, 1989). Specifically, they question whether teachers have sufficient knowledge about the domains that are tested and the tasks they are asked to judge. Also questioned are teachers' abilities to discriminate such constructs as achievement and motivation, and such individual differences as low achievement and specific learning disabilities (Hoge & Butcher, 1984; Salvesen & Undheim, 1994). Another area of concern is the

subjectivity inherent in teachers' judgments (Silverstein, Brownlee, Legutki, & MacMillan, 1983) and the extent to which teachers' expectations and biases may influence student outcomes (Hoge, 1984; Hoge & Butcher, 1984; Sharpley & Edgar, 1986). Given these concerns, and their implications for students, it is reasonable to ask, "How accurate are teachers' judgments of students' performance?"

In this paper, we review literature documenting the accuracy of teachers' judgments of students' academic performance. Our interest in this literature arises from the need to find efficient, cost-effective methods to collect data about students' school achievements for ECLS. The purpose of the ECLS is to further our understanding of student achievement across various content areas (i.e., literacy, mathematical thinking, and general knowledge) and to study the growth of student competence across these areas. Measures that involve teacher judgment are good candidates for accomplishing these goals since they are efficient and cost-effective (Kenny & Chekaluk, 1993). However, before adopting this methodology, it is essential to examine its validity.

Our paper is organized in four sections. First, we examine the methods by which teacher judgments have been evaluated and summarize findings from this research. Then, we examine research findings that relate specifically to the technical adequacy of these measures. Next, we draw implications from this research for using teacher judgment measures for ECLS. Finally, we make recommendations for designing teacher judgment measures that are trustworthy and will adequately serve the purposes of ECLS. In the process of reviewing this literature and writing this paper, we are seeking answers to the following questions:

1) Do findings from previous research support the use of teacher judgments to assess student achievements?

2) In what ways does previous research on teacher judgments address concerns about the validity, reliability, and equity of these assessments?
3) Can previous research findings inform our development of trustworthy measures that rely on teachers' judgments for ECLS?

Our paper focuses on teacher judgments of student academic achievement. Studies concerned with teacher judgments about student social competence, adaptive behaviors, and approaches to learning will not be reviewed here. Those studies are included in another paper that we prepared for ECLS regarding non-cognitive assessments (Meisels, Atkins-Burnett, & Nicholson, 1995).

Evaluating Teachers' Judgments

Methods

By far the most typical way of evaluating the accuracy of teacher judgments is to compare them to student performance on criterion measures with proven validity and reliability—standardized measures of achievement and cognitive abilities (see, for example, Coladarci, 1986; Hoge & Butcher, 1984; Sharpley & Edgar, 1986). However, this theme has many variations. Hoge and Coladarci (1989) suggest three criteria to distinguish measures that rely on teachers' judgments: direct vs. indirect, judgment specificity, and norm-referenced vs. criterion-referenced.

Direct vs. indirect measures. Some researchers use what Hoge and Coladarci (1989) refer to as direct measures (Coladarci, 1986; Hoge & Butcher, 1984; Wright & Wiese, 1988). They present teachers with specific standardized measures, or even specific items from those measures (Coladarci, 1986), and ask them to make judgments about how well particular students will perform on them. For example, Hoge and Butcher (1984) asked teachers to estimate how well their students would perform on the Gates MacGinitie Test of Reading Achievement

that was to be administered in two weeks. Teachers were given a set of cards, each with the name of a student on it, and asked to indicate the grade equivalent score they expected each student to receive on the test. Also, they were asked to rate, on a five-point scale, how confident they were about each prediction. Teachers' judgments were highly predictive of student performance on the achievement test ($r = .85$). However, teachers' ratings of how confident they were about their judgments were not reliably related to residual scores, or to scores based on the difference between predicted and observed scores. This finding suggests that teachers lack confidence in, or are not sensitive to their ability to judge students' performance, even though their judgments are fairly accurate.

In another study (Coladarci, 1986) teachers were asked to make judgments about students' performance on a standardized achievement test at the item level (the SRA Achievement Series). In this study, teachers were interviewed one to two weeks after their students had taken the test. Before the interview, teachers, who were not informed of the results of the testing, divided their students into three groups: 1) performing at grade level, 2) performing one year below grade level, and 3) performing one year above grade level. The interviewer randomly selected two students from each group and asked teachers whether each student responded correctly to specific items on the reading vocabulary, reading comprehension, mathematical concepts, and mathematical computation subtests. On average, teachers evaluated 70-77% of their students' responses correctly. However, teachers' accuracy varied according to differences in students, and across subtests. Specifically, teachers were more accurate in their judgments of high-achieving students and when judging students' performance on test items involving mathematical computation rather than items involving understanding of mathematical concepts. These differences are described in more detail in the

section of this paper that evaluates the criterion-related validity of teacher judgments.

Research designs that incorporate indirect measures are widely used (Kenny & Chekaluk, 1993; Salvesen & Undheim, 1994; Sharpley & Edgar, 1986; Silverstein et al., 1983). Indirect measures consist of procedures in which teachers make judgments about students' overall performance in particular domains (e.g., reading, mathematics, social studies), or about factors associated with achievement (e.g., attitude, effort, general cognitive ability), but do not call for teachers to estimate student performance on specific items adapted from standardized measures. For example, one study asked teachers to rate each child in their class according to the students' current level of achievement in reading vocabulary, reading comprehension, mathematics, general intelligence, and general attitude (Sharpley & Edgar, 1986). The rating scale gave teachers five options that ranged from outstanding to well below average. Teachers' ratings were reliably correlated with students' achievement ($p<.01$).

Another study requested that teachers judge their students' overall performance in oral reading, reading comprehension, spelling, and mathematics (Salvesen & Undheim, 1994). Teachers were given seven options ranging from very poor to very high. Although teachers' ratings were positively skewed, they correlated well with achievement test scores.

Some researchers use indirect measures that present teachers with fine-grained descriptors relating to achievement (Fedoruk & Norman, 1991; Kenny & Chekaluk, 1993; Stevenson, Parker, & Wilkinson, 1976). For example, Fedoruk and Norman (1991) presented teachers with 86 descriptors, compiled from the research literature, that were associated with first grade achievement. Items on this measure included descriptors as specific as "recalls story details", "names colors", "forgets instructions", "never has colds", and "is clumsy and awkward".

Teachers were asked to rate each descriptor on a nine point scale from absolutely contributes to success to absolutely contributes to failure. Their findings suggest that teachers varied considerably in how they interpreted the descriptors.

Finally, some researchers collect both direct and indirect measures of teachers' judgments. Wright and Wiese (1988) asked teachers to rate the achievement and effort of students in their classes in four academic areas: reading, mathematics, language arts, and social studies. In addition, they asked teachers to predict students' national percentile scores in the same subject areas on an upcoming achievement test (the SRA Series, and the Educational Achievement Series). A factor analysis of teachers' ratings indicated that teachers judged students' achievement and effort independently. Regression analyses indicated that teachers' ratings of students' performance in the four academic domains, as well as their estimates of students' performance on the criterion measure, were highly correlated with students' actual performance. Regarding teachers' predictions of students' national percentile scores, standard errors for teachers' estimates amounted to only 4 or 5 percentile points across areas (p. 12).

Hoge and Coladarci (1989) summarize findings from 17 studies that used direct and indirect measures of teacher judgment. They concluded that when direct measures are used to elicit teachers' judgments, they are more accurate than when indirect measures are used. Median correlations were .69 and .62 for direct and indirect measures, respectively. This difference does not appear dramatic, and is not statistically reliable. However, the lowest correlation for the direct measures is substantially higher than the lowest correlation for the indirect measures and the range of correlations from studies using indirect vs. direct measures is larger—.28-.86, in contrast to a range of .48-.92 for studies using direct measures. These differences might be explained by the fact that judgments about performance on particular tests and test items result in more consistent

understandings about the judgment to be made or that direct measures are better matched with the criterion measure than are indirect measures.

Judgment specificity. Researchers vary with regard to the nature and specificity of the judgments they ask teachers to make. Hoge and Coladarci (1989) identify five types of judgments that researchers ask teachers to make, and order them according to the level of specificity the judgment entails: ratings, rankings, grade equivalence, percent correct, and item responses.

Measures that ask teachers to rate general characteristics of their students using a scale (e.g., 1 - 5) have the least specificity in terms of the discriminations they ask teachers to make about students (Hoge & Coladarci, 1989). However, these data can be obtained quickly, with low frustration on the part of teachers, and with reasonable accuracy (Hopkins et al., 1985). Asking teachers to rank their students in terms of their quartile placement in the class (Silverstein et al., 1983), or from most to least capable in reading (Hopkins et al., 1985), yields more specific information, but is more difficult and time consuming for teachers since it requires them to make finer discriminations among students. More specific measures ask teachers to estimate students' grade equivalent scores (Hoge & Butcher, 1984), national percentiles (Wright & Wiese, 1988), or percent correct on specific measures (Coladarci, 1986). The most specific measures ask teachers to consider specific test items and judge whether individual students will answer those items correctly (Coladarci, 1986). According to Coladarci (1986) and Leinhardt (1983), the value of having teachers judge students' performance on particular items is that it reveals teachers' knowledge about what students have or have not mastered in particular domains, and this knowledge should have important implications for the nature and quality of the instruction that students receive.

The distinction between direct and indirect measures also has implications for judgment specificity (Hoge & Coladarci, 1989). Indirect measures differ from direct measures because they are not tied explicitly to a single criterion. For example, researchers who use indirect measures but ask teachers to rank their students from best to worst in particular subject areas (e.g., Hopkins et al., 1985; Luce & Hoge, 1978) are asking teachers to make finer discriminations among students than those who ask teachers to rate students on a fixed scale. Similarly, researchers who use direct measures and request that teachers predict whether individual students will answer particular items correctly (e.g., Coladarci, 1986; Leinhardt, 1983) are asking teachers to make more specific judgments than those who ask teachers to estimate students' grade equivalent scores, or percentile placements, for the test as a whole.

Hoge and Coladarci (1989), in their review of 17 studies that used teacher judgment measures, found that the accuracy of teachers' judgments increased with the specificity of the judgment to be made. Studies that relied on ratings yielded a lower median correlation (.61) than studies relying on rankings (.76), grade equivalents (.70), or item judgments (.70). Also, greater variation was found among studies when ratings and rankings were used (.37 - .92 and .28 - .86, respectively) than when grade equivalents and item judgments were used (.67 - .74 and .67 - .72, respectively). These findings suggest that more specific measures, like more direct measures, yield more consistent judgments because they result in less varied interpretations. However, they argue less against ratings per se than against non-specific ratings.

Norm-referenced vs. criterion-referenced judgments. Some researchers ask teachers to estimate grade equivalent scores or national percentiles. Others ask teachers to rank their students. These activities reflect a norm-referenced approach (Hoge & Coladarci, 1989); that is, students' performance is compared

with the average performance of a specific reference group. Alternatively, when teachers are asked to estimate the number of items individual students will answer correctly on a given measure, or to judge whether students will answer particular test items correctly, this reflects a criterion-referenced approach.

Norm-referenced judgments are more prevalent in the research on teacher judgments than criterion-referenced judgments. However, the proximity of the reference group to the group being judged varies across studies. For example, some researchers ask teachers to compare individual students with other students in their class (Hopkins, George, & Williams, 1985). Others ask teachers to compare students in their class with national samples. This is true for researchers who ask teachers to estimate students' grade equivalent scores (Hoge & Butcher, 1984) or percentile placements on specific measures (Wright & Wiese, 1988). Finally, some researchers ask teachers to compare their current students to all other students they have taught (Wright & Wiese, 1988).

Hoge and Coladarci (1989) claim that whether a judgment is norm-referenced or criterion-referenced does not significantly affect its accuracy. In their review of teacher judgment studies, the median correlation between teachers' judgments and students' performance was .68 for criterion-referenced judgments and .64 for norm-referenced judgments. However, much greater variation was found among studies that asked teachers to make norm-referenced judgments (.28 - .92) than among studies that sought criterion-referenced judgments (.67 - .72). This difference may be due, in part, to the fact that criterion-referenced judgments tend to be more direct and specific. However, it is also possible that the proximity of the reference group affects the accuracy of teachers' judgments; that is, teachers' judgments may be more accurate when their reference group consists of other students in their class rather than when it is a national sample.

Summary. Teacher judgment measures can be characterized in terms of their directness or indirectness, specificity, and norm- or criterion-reference. Direct measures ask teachers to judge students' performance on a particular criterion measure, or on particular items on that measure. Indirect measures ask teachers to make more global judgments; that is, they ask teachers to judge students' overall performance in a particular subject area, or factors associated with achievement (e.g., motivation). Specificity refers to the types of judgments teachers are asked to make (e.g., rating, ranking, etc.). In general, the more direct and specific the judgment, the greater accuracy and consistency obtained. However, it appears that accuracy and consistency can be obtained when teachers are asked to make indirect judgments using a rating scale as long as teachers have correct and consistent understandings about the judgments to be made. Research also suggests that criterion-referenced measures, which ask teachers to compare individual students to a specific standard (e.g., items reflecting a curriculum domain), yield greater consistency than norm-referenced measures. However, it appears that the accuracy of norm-referenced judgments may be related to teachers' familiarity with the reference group (e.g., other students in their class vs. a national sample).

The Technical Adequacy of Teachers' Judgments

Validity

Most researchers focus their evaluation of teacher judgment measures on criterion-related evidence of validity. Within this area of investigation, most researchers provide evidence for concurrent or predictive validity, and some researchers provide evidence related to sensitivity and specificity (Kenny & Chekaluk, 1993; Meisels, Liaw, Dorfman, & Nelson, 1995; Salvesen & Undheim,

1994). Evidence regarding the construct and content validity of these measures is also available.

Criterion-related validity. Studies that evaluate the accuracy of teachers' judgments compared with a specific criterion (e.g., students' performance on standardized tests) yield positive results. Hoge and Caladarci (1989) report a median correlation of .66 between teachers' judgments and students' performance on standardized tests. Similarly, Hoge and Butcher (1984) report a partial regression coefficient of .71 between teachers' judgments and students' actual achievement on standardized tests. Moreover, Coladarci (1986) and Leinhardt (1983), who calculated percentage agreement between teachers' judgments and students' performance on an item by item basis, found that teachers made accurate judgments about whether students would respond correctly about two-thirds of the time (70% and 64% respectively). These findings suggest that teachers are able to make judgments about students performance with a moderate to high degree of accuracy—good news for ECLS as well as practitioners, students, and parents.

However, Hoge and Coladarci (1989) report that judgment accuracy in the studies they reviewed ranged from $r = .28$ to $r = .92$. This finding indicates that teachers' judgments may be more accurate in some circumstances than others, and that accuracy may vary among teachers. For example, researchers who have evaluated the accuracy of teachers' judgments across subject areas have found that teachers' judgments are more accurate in the areas of reading, language arts, and mathematics than in social studies and science (Coladarci, 1986; Hopkins et al., 1985). Furthermore, Coladarci (1986) found that teachers' judgments about students' observable behavior (e.g., math computation) were more accurate than their judgments about activities that are less observable (e.g., reasoning, problem solving). He suggests two possible explanations for this finding. One is that

teachers spend more time on these concrete activities. The other is that, as a result of having students engage in these concrete activities, teachers collect more concrete evidence of students' proficiency in these areas and can use this evidence when making judgments.

Evidence that teachers' judgments can predict students' future performance with reasonable accuracy is also available (Salvesen & Undheim, 1994; Stevenson et al., 1976). For example, Stevenson and his colleagues (1976) followed a group of children from kindergarten through grade 3, asking teachers to rate them five times—twice in kindergarten and once each in grades 1, 2 and 3—on variables relating to cognitive abilities, classroom skills, and personal-social characteristics believed to be important for success in school. Overall, Stevenson et al. found considerable stability in ratings made by kindergarten teachers over a six month period and stability remained high when ratings were made by different teachers two and three years later. Ratings of cognitive abilities and classroom skills were more stable than ratings of students' personal-social behaviors. Also, the predictive value of teachers' ratings for students' academic achievement in reading and mathematics increased from .55 and .50 to .80 and .70, respectively, after kindergarten. These findings echo those that suggest teachers are more consistent in their ratings of children's academic achievement than behavior, and that ratings are more consistent and accurate for older than younger students.

Researchers who include measures of sensitivity and specificity in their evaluation of teacher judgments have typically found satisfactory levels of false positives and false negatives (Kenny & Chekaluk, 1993; Meisels et al., 1995; Salvesen & Undheim, 1994). Sensitivity reflects the extent to which a measure correctly identifies students with characteristics associated with a particular sub-population (e.g., at-risk). Students incorrectly identified by the measure as

belonging to that group can be represented by a ratio that reflects false positive identifications. Conversely, specificity reflects the extent to which a measure correctly excludes students from a sub-population. Students incorrectly excluded from that group can be represented by a ratio that reflects false negative exclusions (Meisels, Henderson, Liaw, Browning, & Ten Have, 1993). Typically, sensitivity and specificity ratios equal to or greater than .80 are highly adequate for assessment purposes.

Kenny and Chekaluk (1993) reported false negative ratios of .23, .12, and .06 for children in kindergarten, grade one, and grade two, respectively. Ratios for false positives were slightly higher, but generally within acceptable limits (.30 for kindergarten, .22 for grade one, and .13 for grade two). Similarly, Meisels et al. (1995) reported high rates of sensitivity and specificity for their developmental checklists. These researchers found that the checklists completed in the fall, which relied on teachers' judgments, were more accurate for predicting students' performance on a standardized measure of achievement administered in the spring (the Woodcock-Johnson Psychoeducational Battery-Revised) than was the standardized measure administered in the fall when compared with itself. Evidence that teachers' judgments have predictive validity is important for ECLS, since we will be relying on teachers' judgments of students' achievement from kindergarten through grade 5.

Construct validity. Silverstein et al. (1983) used a multitrait-multimethod design to examine patterns of relationships among mathematical, reading, and general ability assessed by teacher ratings and standardized achievement tests. Their results indicated higher convergent and discriminant validity for the standardized tests than for teacher ratings, and less method bias (e.g., non-standard administration, uncontrolled subjectivity). Other researchers also report low levels of discriminant validity for teacher judgment measures. Specifically,

teachers' ratings in one achievement area tend to be significantly correlated with those in another (Luce & Hoge, 1978; Pedulla, Airasian, & Madaus, 1980). These findings may indicate bias in teachers' judgments; however, Hoge (1983) suggests that they also may indicate strong relationships among achievement areas. The results of Hoge and Butcher's study (1984) support this theory. They found that teachers' ratings of students' achievement in reading correlated with their ratings of students' general ability and motivation. However, when they compared the magnitudes of the correlations, they found reliable differences, indicating that the ratings were independent. Stevenson et al. (1976) also examined intercorrelations among variables that teachers in their study were asked to rate. They found that teachers were rating individual children differentially on multiple dimensions of behavior.

Some researchers evaluate teachers' judgments in terms of the utility of these judgments in discriminating among groups of students (Hoge, 1983; Kenny & Chekaluk, 1993; Salvesen & Undheim, 1994). Findings across studies indicate that teachers are able to discriminate between students at risk and not at risk, and between low-achieving students and average and high-achieving students. However, one study indicates that teachers may overlook specific disabilities, particularly if students are achieving satisfactorily in school (Salvesen & Undheim, 1994).

Content validity. Evidence for content validity is almost never provided in the literature about teacher judgments. The exception is found in a study by Leinhardt and Seewald (1981), which compared the efficacy of using teachers' judgments to determine "overlap"—the extent to which items on a criterion measure reflect content covered by curriculum materials—with a computer program designed for the same purpose. Their findings indicated that, although both estimates did equally well in predicting students' performance, teachers'

judgments provided a more accurate measure of overlap. According to Leinhardt and Seewald, the computer program tended to underestimate curriculum coverage because it did not account for the instruction teachers provide that is not in curriculum materials. Furthermore, although the computer program may be more objective than teacher judgments, it was much more costly and time consuming.

Summary. ECLS researchers can take comfort in evidence indicating that teachers' judgments have satisfactory criterion-related and predictive validity. However, attention needs to be paid to the domain in which teachers are being asked to make judgments (e.g., reading and math vs. social studies and science) and the characteristics of students targeted by their judgments (e.g., age, ability/achievement level). Concerns have been raised about the extent to which teachers judge such constructs as achievement and motivation independently, and how well they discriminate among groups of students (e.g., low-achievers and students with specific disabilities) that are the targets of their judgments. Research findings about these issues are equivocal. Findings from some studies indicate that teachers have difficulties with such discriminations (Salvesen & Undheim, 1994; Silverstein et al., 1983). Findings from other studies indicate they do not (Hoge & Butcher, 1984). Therefore, it seems important for us, in our development of measures that rely on teachers' judgments, to evaluate ways in which teachers interpret the judgment requests and ensure that they have sufficient knowledge to make appropriate judgments. Finally, the study conducted by Leinhardt and Seewald (1981) suggests that teachers' judgments may provide better evidence of content coverage than more "objective" measures (e.g., analysis of curriculum materials). This provides added justification for ECLS's use of teacher judgment measures to assess students academic development.

Reliability

Reliability has received much less attention than validity in the teacher judgment literature. When addressed, it is evaluated using procedures for test-retest reliability (Fedoruk & Norman, 1991; Hoge, 1983; Stevenson et al., 1976), or inter-rater agreement (Hoge, 1993; Meisels et al., 1995).

Consistency within teachers. Researchers report evidence that suggests teachers' judgments are stable over time. For example, Airasian et al. (1977) provided evidence for the test-retest reliability of teachers' ratings of students' performance in English and mathematics in grade two. Teachers rated their students in each of these areas on two occasions. Ratings were separated by seven months. Correlations between the two ratings were .88 for English and .87 for mathematics. In another study (Stevenson et al., 1976) teachers were asked to rate their students on a variety of variables related to academic success twice in kindergarten and again at the end of grades two and three. Their findings indicate considerable stability between ratings made by kindergarten teachers over six months: correlations ranged from .41 to .74 and stability remained high when the ratings were made by different teachers and separated by two and three years. Finally, Fedoruk and Norman (1991) assessed the consistency of teachers' judgments by requesting that a subset of their sample, 5 of 21, repeat the judgment procedure four to six weeks after the initial administration. They also found a high degree of consistency for individual teachers' judgments—agreement within teachers across two ratings ranged from 81.4% to 98.8%.

Consistency across raters. In contrast to their findings regarding the consistency of individual teachers' judgments, Fedoruk and Norman (1991) found a high degree of variability among teachers in terms of their interpretations of the achievement descriptors they were asked to judge. In their study, teachers were presented with 86 descriptors related to first grade

achievement and asked to indicate the degree to which they thought the descriptor contributed to academic success in grade one. In response to the disparity in teachers' responses, Fedoruk and Norman interviewed the teachers to determine what they were thinking about when they were shown various descriptors. For example, teachers were asked about their thought processes when rating the descriptor "Chinese." Some said they focused on linguistic differences that may place Asian students at a disadvantage. These teachers rated the descriptor as contributing to failure. Other teachers, who focused on the strength of Chinese family units and their motivation to do well, aligned the descriptor with success. Fedoruk and Norman suggest that teachers react differently to different student characteristics, and that the strength of predictive indices is relative to the ecological or situational meaning of those indices.

Wasik and Loven (1980) reviewed research that used observational tools to gather data about students' behavior. Some of their findings shed light on reliability issues relating to measures that rely on teacher judgments in general, and the findings of Fedoruk and Norman (1991) in particular. They identify many factors that influence reliability, but two seem especially relevant here: code complexity and drift. According to Wasik and Loven, the more complex the code, or the greater the number of categories teachers are asked to discriminate, the less consistent are their judgments. Hoge's (1985) review also indicates that measures that include a large number of categories yield less consistent judgments. Related to code complexity is observer drift. According to Wasik and Loven, this occurs when observers, or raters, interpret variables differently or have difficulty keeping categories distinct. This is more likely to occur when measures include a large number of categories. It could be that teachers' judgments in the Fedoruk and Norman study were influenced by both these factors. These findings suggest that researchers should include as small and

discrete a number of categories as possible and train raters well to ensure their understanding of the categories upon which their judgments are based.

Meisels et al. (1995) used zero-order correlations to evaluate evidence for inter-rater reliability between a group of trained raters, and between the raters and teachers who were implementing the Work Sampling System. This is a performance assessment that relies on teachers' judgments in completing developmental checklists, evaluating the contents of student portfolios, and writing summary reports. In their study, high levels of inter-rater reliability between trained raters were achieved (r=.88), compared with moderate levels of inter-rater reliability between raters and untrained teachers (r=.68).

Internal Consistency. Meisels et al. (1995b) also provided evidence for the reliability of their developmental checklist in the form of internal consistency coefficients. This checklist includes items relating to five areas of children's development: art and fine motor, movement and gross motor, concept and number, language and literacy, and personal/social. Alphas for this measure, ranging from .87 to .94, indicate that the checklist is highly reliable.

Summary. Research findings indicate that individual teachers' judgments are consistent across time. However, evidence suggests that judgments across teachers may vary depending on the number of categories teachers are asked to discriminate, the training the teachers receive, and the extent to which the judgment has differential interpretations. These findings emphasize the need in ECLS to evaluate teachers' independent understandings of, and ability to make, targeted judgments. They also suggest that researchers should provide as much training to teachers/raters as possible in order to increase consistency. If extensive training for participants is not feasible, as is often the case for large national studies, researchers can avoid some of the pitfalls mentioned above through extensive piloting of instruments. ECLS is conducting such pilots.

Equity

Research on equity in teacher judgments has focused on issues of gender and ability. Specifically, researchers have sought evidence to determine whether teachers rate girls more highly than boys, and whether teachers' judgments favor students who have high abilities (Hoge, 1984; Hoge & Butcher, 1984; Hoge & Coladarci, 1989; Sharpley & Edgar, 1986; Stevenson et al., 1976). Generally, very little support for these biases has been reported. When evidence is found that implicates bias in teachers' judgments, it typically suggests that teachers rate all of their students positively (Salvesen & Undheim, 1994; Sharpley & Edgar, 1986; Silverstein et al., 1983). However, even with this skew, teacher assessments correlate well with achievement tests (Salvesen & Undheim, 1994).

Assumptions that teachers are biased in their evaluations of students are not new. In 1968 Rosenthal and Jacobson generated the "teacher expectancy hypothesis"—that students' performance in school can be attributed, at least in part, to teachers' expectations of them. Since then, Hoge (1984) has challenged this hypothesis. He examined studies that used induction techniques, providing teachers with names of students expected to do well or poorly, or with contrived information about test scores, as well as more naturalistic measures that related teachers' existing expectations to students' actual classroom behavior or performance. While admitting that there are problems with how teacher expectancy has been conceptualized and measured by researchers, Hoge concluded that there is no convincing evidence in the literature that teachers' judgments are affected by inductive techniques, and only weak and inconsistent support relating teachers' own expectations to behavioral and academic outcomes for students. For example, Hoge (1984) cites researchers who used induction techniques to manipulate teachers' expectations for students, and then interviewed them at the end of their investigations (Fleming & Anttonen, 1971;

Jose & Cody, 1971; Mendels & Flanders, 1973). In each of these investigations, teachers confided that, "for one reason or another, they had discounted the information" that might have biased their judgments (Hoge, 1984, p. 218).

Similarly, in studies where teachers' ratings seemed to favor girls, or high ability students, ratings were found to reflect students' performances on standardized achievement tests. For example, Sharpley and Edgar (1986) found that teachers assigned lower ratings to grade three boys on three academic variables: vocabulary, comprehension, and mathematics. However, performance data in this study supported their ratings. Furthermore, teachers' ratings of attitude were not reliably correlated with teachers' ratings of other variables, suggesting that teachers' evaluations of attitude did not result in lower ratings on academic variables.

Hoge and Butcher (1984) examined teachers' judgments for evidence of both gender and ability bias, first as a group and then on an individual basis. Their findings indicate that gender does not reliably affect teachers' judgments. However, they did find that some teachers (3/12) overestimated the achievement of high-ability students. Also, Hoge and Butcher report correlations between teachers' judgments of students' achievement and motivation. This finding may suggest a bias in favor of high ability or highly motivated students. However, the relationship between achievement and motivation is well documented in other literatures.

Although research about teachers' judgments does not provide clear evidence for gender or ability bias, several studies indicate that teachers are better judges of students with high ability than students with low ability, or specific learning difficulties (e.g., dyslexia, mental handicaps), and better at judging students as their years in school increase (Coladarci, 1986; Hoge & Butcher, 1984; Kenny & Chekaluk, 1993). Coladarci (1986) suggests it is easier to

be accurate when judging high achieving students because teachers can work with a "general response set" that is effective for high achieving students (p. 145). These students are likely to do well on a test, and will more likely than not answer items correctly. Therefore, if a teacher judges that a student will answer all items on a test correctly, and the student answers 80% correctly, the teacher achieves a relatively high degree of accuracy, regardless of how confident he or she may be about the student's ability to respond to particular questions. In contrast, low-achieving students are not likely to do well, but it is difficult to pinpoint how many and which specific errors will occur. Furthermore, students with learning difficulties, especially learning disabilities, tend to be erratic in their performance and this makes it difficult to predict their performance on a given day or a particular task. The same is true for young children. As children spend more time in school, their performance is more easily assessed (Kenny & Chekaluk, 1994; Sharpley & Edgar, 1986).

Somewhat contrasting findings about bias are presented in a study by Bennet, Gottesman, Rock, and Cerullo (1993). This study was designed to determine whether students' gender and/or behavior influenced teachers' judgments of their academic skill. The researchers proposed a path model in which gender and academic skill were hypothesized to influence teachers' judgments of academic skill directly, and indirectly through perceptions of students' behavior. Students were 794 children in kindergarten and grades 1 and 2. Males constituted 52% of the sample. Instruments included a standardized measure of students' academic skill, behavior ratings, and teachers' academic judgments.

Regarding the hypothesized relationships in the proposed model, the researchers claim that behavior influenced teachers' judgments of kindergarten students' academic skill. Also, they claim that gender influenced teachers' judgments of behavior and, subsequently, behavior influenced teachers'

judgments of grade 1 and 2 students' academic skill. Some problems with this study may, however, call these conclusions into question.

First, the proposed model assumes unidirectional relations among variables. Thus, the researchers, while acknowledging the possibility of alternative explanations do not probe them statistically. Second, the researchers did not use the classroom as the unit of analysis, and the ranges of relationships indicate differences among classes/teachers (a .30 spread for most variables). This is consistent with other researchers' findings, and one of the reasons that Hoge and Coladarci (1989) emphasize the need to use the classroom as the unit of analysis, at least during initial analyses. When data are analyzed on a class by class basis, reliable differences have been detected. For example, Hoge and Butcher (1984) found that three of twelve teachers in their study were overestimating the performance of high-ability students and underestimating the ability of low-ability students. Hopkins et al. (1985) suggest that pooling data across classrooms ignores differences among teachers that may cause relationships between teacher judgments and criterion measures to be under or over-estimated.

A third problem with the conclusions of this study concerns the magnitude of their findings for gender. Using their formula (indirect relation = the product of the direct relations), we obtain an indirect relationship for gender of .10 and .08 with academic grades and academic ratings, respectively. This relation is weak in comparison to the .46 and .49 correlations between tested academic skill and academic grades, and tested academic skill and academic ratings, respectively, suggesting that teachers rely more on their knowledge of students' academic skill when making judgments of academic performance than on gender. This is consistent with other findings in our review of the literature and supportive of the use of teacher judgments in ECLS.

Finally, the researchers report that behavior ratings are linked with teachers' judgments regarding students' academic performance in all grades (K-2). Is this the result of bias on the part of teachers, or is it possible that students who are well behaved tend to do well academically, and/or that doing well academically reinforces or promotes "good" behaviors on the part of students? These researchers also find that girls at this level (K through 2) are consistently given higher behavior ratings than boys. Again, this could reflect bias on the part of teachers or the fact that the criteria against which good behavior in school is judged is more consistent with girls' behavior, especially at this age. These issues are not addressed clearly in this study, or in the other studies we reviewed, resulting in equivocal interpretations. Overall, these findings indicate that, while some teachers' judgments may reflect bias of one sort or another, teachers as a group base their judgments of students' academic performance on their knowledge of students' academic skill.

Although the teacher judgment literature focuses primarily on potential bias consequent upon gender and ability differences, possible bias related to linguistic and cultural diversity among students requires investigation as well (Darling-Hammond, 1993; Garcia & Pearson, 1993; Madaus, 1994). Teachers may make inaccurate judgments about students when they lack knowledge of cultural and linguistic differences among students.

There are few studies in the teacher judgment literature that explicitly and systematically examine teachers' judgments for evidence of linguistic and cultural bias. One recent study, which examined the reliability and validity of portfolio assessments (LeMahieu, Eresh, & Wallace, 1993-94), mentioned that ratings of students' work had been examined for presence of cultural or racial bias, and none was found. Another study (Bahr, Fuchs, Stecker, & Fuchs, 1991) examined the extent to which teachers' judgments about black and white

"difficult-to-teach" boys were racially motivated. The teachers in their study judged significantly more black males as appropriate for referral to special education. However, achievement data from that study support the teachers' judgments.

A third study (Ross & Jackson, 1991) asked teachers to make three judgments based on their reading of 12 case studies (six concerning black male students and six concerning black female students). These judgments involved a) predicting the student's success in the current school year, b) predicting future success and c) rating the desirability of having that student in their classroom. Each case description included information about students' academic performance (e.g., standardized test scores) and behavior. Teachers' expectations improved as achievement improved for both current year and future success predictions and they did not differentiate between submissive and nonsubmissive students when predicting academic success. However, teachers' expectations regarding current end-of-school-year averages were slightly higher for the girls presented in the case studies, indicating a preference for high achieving, female and submissive students. Black male non-submissive students were least preferred by the teachers in this study. Unfortunately, the researchers did not include cases that involved white male non-submissive students, so our ability to conclude that teachers' preferences were associated with racial bias is severely limited.

Garcia and Pearson (1991) describe ways that teachers' judgments and tasks that rely on teachers' judgments can be biased against students with minority status. They point out that all measures reflect the values, norms, and mores of those who develop and use them. However, they note that equity can be achieved when examiners have knowledge of cultural and linguistic differences among students, and are willing to consider the backgrounds of individual students and seek a range of possible explanations for students' behaviors.

Summary. In most of the studies we reviewed, support for gender and ability bias was not found for teachers as a group. However, findings suggest that some teachers may have biases that influence their judgments, and that teachers as a group find it more difficult to judge students whose performances are inconsistent (e.g., young children and students with learning difficulties). Additional research is needed to explore the potential impact of linguistic and cultural differences between students and teachers on teachers' judgments, since empirical data regarding these issues is scant. In general, the findings regarding bias support our moving forward with teacher judgment measures in the ECLS.

Limitations of Research on Teachers' Judgments

Before drawing implications from the teacher judgment literature and making recommendations for developing meaningful and trustworthy measures, it is important to acknowledge the limitations in this literature. These limitations can be used as benchmarks for developing new measures that rely on teacher judgment.

First, considerable variation exists among the teacher judgment measures in terms of directness, judgment specificity, and reference group. Even within each of these categories there is substantial variation. For example, an indirect measure can be a judgment about a global construct such as reading achievement, or a more specific aspect of that construct such as knowledge of letter-sound correspondence. This makes it difficult to draw conclusions across studies.

Second, researchers' descriptions of the judgments that teachers in their studies are asked to make lack detail. This makes it difficult to either replicate their measures, or determine ways in which their measures can be improved.

Third, care is not always taken to ensure that the criterion measure against which teachers' judgments are evaluated is appropriate for the judgments being made, or well understood by those making the judgments. Often, criterion measures are poorly aligned with classroom curricula and the kinds of judgments teachers make on a daily basis, and with constructs about which teachers are being asked to make judgments.

For example, several researchers in the studies reviewed in this paper used the Peabody Picture Vocabulary Test (PPVT) as the criterion measure for assessing the accuracy of teachers' judgments about students' general ability (Kenny & Chekaluk, 1993; Sharpley & Edgar, 1986). This is an inappropriate use of this tool, since general ability entails much more than is measured by the PPVT. Furthermore, this measure is not a fair indicator of the general abilities of students from culturally and linguistically diverse populations, or students who have language disabilities (Washington & Craig, 1992). Discussion of biases in many of the criterion measures used in studies of teachers' judgments is conspicuous by its absence.

Fourth, it is striking that the onus is on teachers to match standardized measures, particularly in terms of identifying students at risk, or distinguishing between low-achievers and students with specific learning disabilities. Again, there is little mention of the criterion measures' limitations, or the fact that teachers, because they observe and interact with their students on a daily basis, may be in the best position to make judgments about them.

Fifth, there is a lack of standardization concerning the form and administration of teacher judgment measures across studies of teachers' judgments. According to Hoge (1983; 1984), variation is ignored and all measures are treated as equal. Also, Hoge points out a tendency, on the part of researchers, to develop unique measures for their studies, rather than using established

measures that have proven their technical adequacy. Meisels, Dorfman, and Steele (1994) suggest that assessments that are not standardized may not be trustworthy or fair. They remind us that "standardization" does not necessarily imply standard scores, or norm-referenced assessment. Rather, standardization refers to "formal rules of operation and explicit principles of interpretation [that have] been studied sufficiently to understand how different groups of children in different situations will react to a particular assessment" (p. 204).

Finally, most studies pool data across classrooms to form an undifferentiated group. This ignores differences among teachers and between classrooms and may result in misinterpretations of findings.

Summary and Implications

We return now to the questions raised at the outset of this paper. First, do the findings of previous research justify using teacher judgments to assess students' achievements? Overall, the research indicates that teachers can make accurate judgments of students' academic performance. Correlations between their judgments and more standardized, objective measures of achievement have been as high as .80 or .90. Furthermore, some investigators have shown teachers' judgments to be even better predictors of students' future performance than standardized measures (Meisels et al., 1995).

In addition to the above criterion-related evidence favoring teachers' judgments, evidence for construct and content validity is available. Although some results suggest that teachers have difficulty discriminating between such constructs as achievement and motivation, and among diverse groups of students, researchers who have probed these results report that teachers judge these constructs independently, and are able to identify students having difficulty in school. Also, there is evidence that measures that rely on teachers'

judgments may have greater content validity than more standardized, objective measures (Leinhardt & Seewald, 1981) since they are more closely tied to the classroom curriculum. Finally, evidence concerning the reliability of teachers' judgments is available, as well as evidence that suggests that teachers are fair in their judgments about students, at least with respect to gender and behavioral correlates of achievement (e.g., students' attitudes).

These findings support ECLS use of measures that rely on teachers' judgments to gather data about students as they progress from kindergarten through grade five. For research purposes, the degree of accuracy achieved through teacher ratings (for studies reviewed in Hoge and Coladarci, the median r for ratings was .61, with a maximum r of .92) is entirely appropriate. Furthermore, these measures have proven to be relatively inexpensive and unobtrusive, and the least time-consuming and cognitively demanding judgments for teachers to make.

Also, these findings respond to concerns raised about the validity, reliability, and equity of measures that rely on teachers' judgments, the second question raised in our introduction. The fact that evidence suggests that teachers' judgments are valid and reliable, and that evidence supporting the teacher expectancy hypothesis and other forms of teacher bias is weak, is encouraging for researchers and practitioners who are developing and implementing such assessments.

The third and final question in our introduction asked how previous research findings can inform our development of trustworthy measures that rely on teachers' judgments. In answering this question, it is important to consider research findings that suggest that teachers' judgments can be more or less accurate, depending on a variety of factors. For example, teachers' judgments tend to be affected by the directness of the measure and the specificity of the

judgment they are asked to make. Also, teachers tend to be better at judging the performance of high achieving students than students who have difficulties in school, particularly students with specific learning disabilities, and their judgments are more reliable as students progress through school. Finally, teachers' judgments are affected by what Wasik and Loven (1980) refer to as code specificity and observer drift. Therefore, Hoge and Coladarci (1989) suggest that, although many existing measures that rely on teachers' judgments are valid, they can still be improved. It is in light of this suggestion that we make the following recommendations for the development of teacher judgment measures for ECLS.

Recommendations

The following recommendations, informed by the teacher judgment literature, are meant to guide researchers in general and ECLS in particular in developing meaningful and trustworthy teacher judgment measures. However, we recognize that researchers must make their final decisions regarding instrument development in accordance with the purposes and constraints of their studies.

- In general, it appears that more direct measures yield more accurate and meaningful judgments about students' academic achievements. This is likely because direct measures present for teachers a very clear criterion against which to judge students' performance. However, our reading of the literature indicates indirect measures can also provide accurate and meaningful information about students' performance, as long as the items on the teacher judgment scale are highly specific, teachers are presented with as small a number of clear and distinct categories as possible, and researchers ensure teachers' understanding of those scale items and categories.

- It also appears that judgments that are criterion-referenced and require a high degree of specificity yield more accurate and meaningful information about students. The literature points out that ratings are less specific than rankings or assigned scores. However, this does not mean that ratings yield unsatisfactory levels of accuracy. We would emphasize that specificity is a relative construct and that, for research purposes, ratings yield adequate levels of accuracy (.60 and higher) and are much more efficient and "user friendly" than the alternatives. These are important considerations for ECLS because we will be asking a large number of teachers to contribute their valuable time to completing our measures.
- The previous two recommendations point to the need for teachers to be knowledgeable about the constructs and students they are being asked to judge, as well as the judgment task. The ideal approach to meeting this need involves extensively training teachers who participate in studies, and probing their interpretations of the judgment tasks to avoid misunderstandings. However, when logistic features of the study prohibit this, as in the case of ECLS, extensive piloting of the instruments can address similar issues. In any case, it is very important to evaluate the reliability of teachers' judgments.
- Researchers can evaluate the trustworthiness of the data they gather from teachers' judgments using a multitrait-multimethod design. In the case of ECLS, this is accomplished by collecting data regarding a variety of student characteristics (i.e., social-emotional and behavioral development in addition to academic achievement) from multiple sources (i.e., parent reports and cognitive assessments, in addition to teacher judgments).
- Teachers' judgments can reflect deep knowledge about students that is constructed over time and across many contexts. Nevertheless, all users of

teacher judgment measures should be sensitive to the inherent subjectivity of these measures. Despite the absence of strong evidence for teacher bias in the literature we reviewed, it must be evaluated carefully.

- Finally, it is important to pay close attention to issues related to the technical adequacy and standardized administration of teacher judgments in order to ensure that assessments are valid, reliable, and fair. In particular, students' backgrounds and cultural contexts may interact with assessment outcomes.

References

Bahr, M. W., Fuchs, D., Stecker, P. M., & Fuchs, L. S. (1991). Are teachers' perceptions of difficult-to-teach students racially biased? School Psychology Review, 20, 599-608.

Bennett, R. E., Gottesman, R. L., Rock, D. A., and Cerullo, F. (1993). Influence of behavior perceptions and gender on teachers' judgments of students' academic skill. Journal of Educational Psychology, 85, 347-356.

Calfee, R., & Hiebert, E. (1991). Teacher assessment of achievement. Advances in Program Evaluation, Vol. 1A (pp. 103-131). JAI Press.

Coladarci, R. (1986). Accuracy of teacher judgments of student responses to standardized test items. Journal of Educational Psychology, 78, 141-146.

Darling-Hammond, D. (1993, March). Equity issues in performance-based assessment. Paper presented at The Symposium on Equity and Educational Testing and Assessment, Washington, DC.

Fedoruk, G. M., & Norman, C. A. (1991). Kindergarten screening predictive inaccuracy: First-grade teacher variability. Exceptional Children, 57, 258-263.

Fleming, E., & Anttonen, R. (1971). Teacher expectancy or My Fair Lady. American Educational Research Journal, 8, 241-252.

Garcia, E. G., & Pearson, P. D. (1991). The role of assessment in a diverse society. In E. H. Hiebert (Ed.), Literacy for a diverse society (pp. 253-278). New York: Teachers' College Press.

Garcia, G. E., & Pearson, D. P. (1993). Assessment and diversity. Unpublished manuscript.

Hoge, R. D. (1983). Psychometric properties of teacher-judgment measures of pupil aptitudes, classroom behaviors, and achievement levels. Journal of Special Education, 17, 401-429.

Hoge, R. D. (1984). The definition and measurement of teacher expectations: Problems and prospects. Canadian Journal of Education, 9, 213-228.

Hoge, R. D., & Butcher, R. (1984). Analysis of teacher judgments of pupil achievement levels. Journal of Educational Psychology, 76, 777-781.

Hoge, R. D., & Coladarci, T. (1989). Teacher-based judgments of academic achievement: A review of the literature. Review of Educational Research, 59, 297-313.

Hopkins, K. D., George, C. A., & Williams, D. D. (1985). The concurrent validity of standardized achievement tests by content area using teachers' ratings as criteria. Journal of Educational Measurement, 22, 177-182.

Jose, J., & Cody, J. (1971). Teacher-pupil interaction as it relates to attempted changes in teacher expectancy of academic ability and achievement. American Educational Research Journal, 8, 39-50.

Jussim, L., & Eccles, J. S. (1992). Teacher expectations II: Construction and reflection of student achievement. Journal of Personality and Social Psychology, 63, 947-961.

Kenny, D. T., & Chekaluk, E. (1993). Early reading performance: A comparison of teacher-based and test-based assessments. Journal of Learning Disabilities, 26, 227-236.

Leinhardt, G. (1983). Novice and expert knowledge of individual students' achievement. Educational Psychologist, 18, 165-179.

Leinhardt, G., & Seewald, A. M. (1981). Overlap: What's tested, what's taught? Journal of Educational Measurement, 18, 85-96.

LeMahieu, P. G., Eresh, J. T., & Wallace, R. C. (1993-94). Using student portfolios for public accounting. The School Administrator. American Association of School Administrators.

Luce, S. R. & Hoge, R. D. (1978). Relations among teacher rankings, pupil-teacher interactions, and academic achievement: A test of the teacher-expectation hypothesis. American Educational Research Journal, 15, 489-500.

Madaus, G. F. (1994). A technological and historical consideration of equity issues associated with proposals to change our nation's testing policy. In M. T. Nettles & A. L. Nettles (Eds.), Equity in educational assessment and testing (pp. 23 -68). Boston: Kluwer Academic Publishers.

Meisels, S. J., Atkins-Burnett, S., & Nicholson, J. (1995). Assessment of social competence, adaptive behaviors, and approaches to learning with young children. Ann Arbor, MI: University of Michigan Assessment Projects.

Meisels, S. J., Dorfman, A., & Steele, D. (1994). Equity and excellence in group-administered and performance-based assessment. In M. T. Nettles & A. L. Nettles (Eds.), Equity in educational assessment and testing (pp. 243-264). Boston: Kluwer Academic Publishers.

Meisels, S. J., Henderson, L. W., Liaw, F.-R., Browning, K., & Ten Have, T. (1993). New evidence for the effectiveness of the Early Screening Inventory. Early Childhood Research Quarterly, 8, 327-346.

Meisels, S. J., Liaw, F., & Dorfman, A., & Nelson, R. (1995). The Work Sampling System: reliability and validity of a performance assessment for young children. Early Childhood Research Quarterly, 10, 277-296.

Mendels, G., & Flanders, J. P. (1973). Teachers' expectations and pupils' performance. American Educational Research Journal, 10, 203-212.

National Opinion Research Center. (1993). Early childhood longitudinal study: Kindergarten cohort. (Report No. RFP93-027). Chicago: Author.

Pedulla, J. J., Airasian, P. W., & Madaus, G. F. (1980). Do teacher ratings and standardized tests yield the same information? American Educational Research Journal, 17, 303-307.

Rosenthal, R. & Jacobson, L. (1968). Pygmalion in the classroom: Teachers' expectations and students' intellectual development. New York: Holt, Rinehart, & Winston.

Ross, S. I., & Jackson, J. M. (1991). Teachers' expectations for Black males' and Black females' academic achievement. Personality and Social Psychology Bulletin, 17, 78-82.

Salvesen, K. A., & Undheim, J. O. (1994). Screening for learning disabilities. Journal of Learning Disabilities, 27, 60-66.

Sharpley, C. F., & Edgar, E. (1986). Teachers' ratings vs. standardized tests: An empirical investigation of agreement between two indices of achievement. Psychology in the Schools, 23, 106-111.

Silverstein, A. B., Brownlee, L., Legutki, G., & MacMillan, D. L. (1983). Convergent and discriminant validation of two methods of assessing three academic traits. Journal of Special Education, 17, 63-68.

Stevenson, H. W., Parker, T., Wilkinson, A., Hegion, A, & Fish, E. (1976). Predictive value of teachers' ratings of young children. Journal of Educational Psychology, 68, 507-517.

Stiggins, R. J. (1987). Design and development of performance assessments. Educational Measurement: Issues and Practices, Fall, 33-42.

Taylor, C. (1994). Assessment for measurement or standards: The peril and promise of large-scale assessment reform. American Educational Research Journal, 31, 231-262.

Washington, J. A., & Craig, H. K. (1992). Performances of low-income African-American preschool and kindergarten children on the Peabody Picture Vocabulary Test-Revised. Language, Speech, and Hearing Services in Schools, 23, 329-333.

Wright, D., & Wiese, M. J. (1988). Teacher judgment in student evaluation: A comparison of grading methods. Journal of Educational Measurement, 82, 10-14.

Listing of NCES Working Papers to Date

Number	Title	Contact
94-01	Schools and Staffing Survey (SASS) Papers Presented at Meetings of the American Statistical Association	Dan Kasprzyk
94-02	Generalized Variance Estimate for Schools and Staffing Survey (SASS)	Dan Kasprzyk
94-03	1991 Schools and Staffing Survey (SASS) Reinterview Response Variance Report	Dan Kasprzyk
94-04	The Accuracy of Teachers' Self-reports on their Postsecondary Education: Teacher Transcript Study, Schools and Staffing Survey	Dan Kasprzyk
94-05	Cost-of-Education Differentials Across the States	William Fowler
94-06	Six Papers on Teachers from the 1990-91 SASS and Other Related Surveys	Dan Kasprzyk
94-07	Data Comparability and Public Policy: New Interest in Public Library Data Papers Presented at Meetings of the American Statistical Association	Carrol Kindel
95-01	Schools and Staffing Survey: 1994 papers presented at the 1994 Meeting of the American Statistical Association	Dan Kasprzyk
95-02	QED Estimates of the 1990-91 Schools and Staffing Survey: Deriving and Comparing QED School Estimates with CCD Estimates	Dan Kasprzyk
95-03	Schools and Staffing Survey: 1990-91 SASS Cross-Questionnaire Analysis	Dan Kasprzyk

Listing of NCES Working Papers to Date (Continued)

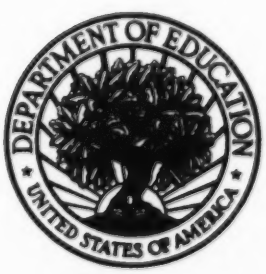
Number	Title	Contact
95-04	National Education Longitudinal Study of 1988: Second Follow-up Questionnaire Content Areas and Research Issues	Jeffrey Owings
95-05	National Education Longitudinal Study of 1988: Conducting Trend Analyses of NLS-72, HS&B, and NELS:88 Seniors	Jeffrey Owings
95-06	National Education Longitudinal Study of 1988: Conducting Cross-Cohort Comparisons Using HS&B, NAEP, and NELS:88 Academic Transcript Data	Jeffrey Owings
95-07	National Education Longitudinal Study of 1988: Conducting Trend Analyses HS&B and NELS:88 Sophomore Cohort Dropouts	Jeffrey Owings
95-08	CCD Adjustments to the 1990-91 SASS: A Comparison of Estimates	Dan Kasprzyk
95-09	The Results of the 1993 Teacher List Validation Study (TLVS)	Dan Kasprzyk
95-10	The Results of the 1991-92 Teacher Follow-up Survey (TFS) Reinterview and Extensive Reconciliation	Dan Kasprzyk
95-11	Measuring Instruction, Curriculum Content, and Instructional Resources: The Status of Recent Work	Sharon Bobbitt & John Ralph
95-12	Rural Education Data User's Guide	Samuel Peng

Listing of NCES Working Papers to Date (Continued)

Number	Title	Contact
95-13	Assessing Students with Disabilities and Limited English Proficiency	James Houser
95-14	Empirical Evaluation of Social, Psychological, & Educational Construct Variables Used in NCES Surveys	Samuel Peng
95-15	Classroom Instructional Processes: A Review of Existing Measurement Approaches and Their Applicability for the Teacher Follow-up Survey	Sharon Bobbitt
95-16	Intersurvey Consistency in NCES Private School Surveys	Steven Kaufman
95-17	Estimates of Expenditures for Private K-12 Schools	Steve Broughman
95-18	An Agenda for Research on Teachers and Schools: Revisiting NCES' Schools and Staffing Survey	Dan Kasprzyk
96-01	Methodological Issues in the Study of Teachers' Careers: Critical Features of a Truly Longitudinal Study	Dan Kasprzyk
96-02	Schools and Staffing Survey (SASS): 1995 Selected papers presented at the 1995 Meeting of the American Statistical Association	Dan Kasprzyk
96-03	National Education Longitudinal Study of 1988 (NELS:88) Research Framework and Issues	Jeffrey Owings
96-04	Census Mapping Project/School District Data Book	Tai Phan

Listing of NCES Working Papers to Date (Continued)

Number	Title	Contact
96-05	Cognitive Research on the Teacher Listing Form for the Schools and Staffing Survey	Dan Kasprzyk
96-06	The Schools and Staffing Survey (SASS) for 1998-99: Design Recommendations to Inform Broad Education Policy	Dan Kasprzyk
96-07	Should SASS Measure Instructional Processes and Teacher Effectiveness?	Dan Kasprzyk
96-08	How Accurate are Teacher Judgments of Students' Academic Performance?	Jerry West


DEPARTMENT OF EDUCATION
UNITED STATES OF AMERICA